Computational Biology and Chemistry 66 (2017) 57–62

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

PrAS: Prediction of amidation sites using multiple feature extraction

Tong Wang[a], Wei Zheng[a], Qiqige Wuyun[a], Zhenfeng Wu[a], Jishou Ruan[a,b], Gang Hu[a], Jianzhao Gao[a,*]

[a] School of Mathematical Sciences and LPMC, Nankai University, Tianjin, 300071, China
[b] State Key Laboratory of Medicinal Chemical Biology, Nankai University, Tianjin, 300071, China

ARTICLE INFO

Article history:
Received 30 July 2016
Received in revised form 27 November 2016
Accepted 28 November 2016
Available online 29 November 2016

Keywords:
Posttranslational modification (PTM)
Amidation sites
Support vector machine (SVM)
Positive contribution feature selection

ABSTRACT

Amidation plays an important role in a variety of pathological processes and serious diseases like neural dysfunction and hypertension. However, identification of protein amidation sites through traditional experimental methods is time consuming and expensive. In this paper, we proposed a novel predictor for Prediction of Amidation Sites (PrAS), which is the first software package for academic users. The method incorporated four representative feature types, which are position-based features, physicochemical and biochemical properties features, predicted structure-based features and evolutionary information features. A novel feature selection method, positive contribution feature selection was proposed to optimize features. PrAS achieved AUC of 0.96, accuracy of 92.1%, sensitivity of 81.2%, specificity of 94.9% and MCC of 0.76 on the independent test set. PrAS is freely available at https://sourceforge.net/p/praspkg.

1. Introduction

Many bioactive peptides must be amidated to exhibit full activity. In peptide hormones, amidation is a common posttranslational modification (PTM) that is mediated in a two-step process by the hydroxylase and lyase activities of the bifunctional enzyme, peptidylglycine alpha-amidation monooxygenase (Driscoll et al., 1999). Amidation had also been reported involved in a variety of pathological processes such as neural dysfunction (Bousquet-Moore et al., 2010), sleep apnea (Kuyama et al., 2009), cancer (Dennison et al., 2009; Rocchi et al., 2001), and hypertension (Shimosawa et al., 2000). Previous reports on the interaction of amidated peptides with their corresponding receptors suggested the activity of amidation to be a crucial determinant of ligand-receptor interaction (Bradbury and Smyth, 1991; Edisom et al., 1999; Walsh and Jefferis, 2006). Actually, amidation is linked to preventing ionization of the peptide's C-terminus and could render it more hydrophobic and thus potentially better able to bind its receptor (Walsh and Jefferis, 2006; Bignon et al., 1998). As for the distribution of amidation activity, subsequent studies have also demonstrated that amidation activity is widely distributed – it is present in almost every tissue including heart, thyroid, hypothalamus, adrenal medulla, submandibular gland, pancreas, intestine, bone, prostate and seminal fluid, even in serum (Bradbury and Smyth, 1991).

One of the key points in studying amidation is to determine amidated proteins and the corresponding sites. Current experimental methods such as mass spectrometry (Kuyama et al., 2009), radioimmunoassay (RIA) and immunoprecipitation (Yamaguchi et al., 2007), are often labor-intensive, time-consuming and expensive. Therefore, computational methods are required to complement experimental methods and to better understand amidation. Although amidation sites are the fourth most among all PTM types calculated by the PTM Statistics Curator web server (Khoury et al., 2011), there isn't any avaliable software package or website for amidation sites so far.

To address this issue, we proposed a computational method to Predict Amidation Sites (PrAS). To extract as much information as possible, we incorporated four feature types, consisting of position-based features, physicochemical and biochemical properties features, predicted structure-based features and evolutionary information features. In order to remove the redundancy of the features, an original feature selection method, positive contribution feature selection (PCFS), was further to optimize the features based on support vector machine (SVM) classifier. We also analyzed the optimized features from the biochemical background. The proposed method achieved AUC of 0.96, accuracy of 92.1%, sensitivity

* Corresponding author.
E-mail addresses: tongwang1992@mail.nankai.edu.cn (T. Wang), jlspzw139@sina.com (W. Zheng), wuyunqqg@163.com (Q. Wuyun), wu_zhenfeng@mail.nankai.edu.cn (Z. Wu), jsruan@nankai.edu.cn (J. Ruan), huggs@nankai.edu.cn (G. Hu), gaojz@nankai.edu.cn (J. Gao).

http://dx.doi.org/10.1016/j.compbiolchem.2016.11.004
1476-9271/

of 81.2%, specificity of 94.9% and MCC of 0.76, in the independent test set. Most importantly, PrAS fills the blank of the research in this area and offers convenience for future studies such as ligand-receptor interaction, function of neurotransmitter, and relevant pathological processes.

2. Materials and methods

2.1. Datasets

All amidation data with experimentally validated were downloaded from Uniprot (Version: March 2015). A total of 692 protein sequences with amidation sites located in internal regions were obtained, because C-terminal sites in sequences actually are seen as the reaction products of amidation (Kolhekar et al., 1997; Cui et al., 2013). For the next step, on one hand, a majority of mammalian and insect peptide hormones possess alpha-amide moiety (Merkler, 1994; Eipper and Mains, 1988) that arises from the oxidative cleavage of Glycine-extended prohormones (Bradbury et al., 1982). That indicates the amide group that resembles 'XGXX' contributes most for amidation, where 'X' means any amino acid and 'G' means Glycine (Wilkins et al., 1999). On the other hand, when calculating without this limitation of mode 'XGXX' (Cui et al., 2013), the identified site is strongly relevant to the amino acid that follows behind it. For these reasons, we included only amino acid sties followed by 'G' to improve the reliability by removal of query sites that had extremely high correlation to the amino acid immediately following it. After this step, 488 protein sequences remained. In order to extract sequence information as conveniently as possible, we removed the sequences that contain less than 21 amino acids as previous research did (Cui et al., 2013), and thus obtained 416 sequences. If the sequences in the dataset are highly homologous, the accuracy of prediction can be overfitted. In order to reduce the homology, we clustered the protein sequences with a threshold of 30% identity using BLASTClust (Alva et al., 2016). All annotated amidation sites are regarded as positive samples, while non-annotated amidation sites, which follows 'XGXX' mode, are regarded as negative samples. One 21-residue peptide fragment for each annotated/non-annotated amidation site was extracted. Ten amino acids are at each side of amidation site. Finally, the dataset contains 209 sequences with 497 amidation sites and 1834 non-amidation sites.

We randomly selected 139 sequences as training set including 332 positive sites and 1202 negative sites, and the remaining 70 sequences as independent test set including 165 positive sites and 632 negative sites. As we can see, the ratio of positive to negative samples was about 1:4 for both training set and independent test set. Since we have removed redundancy of dataset using BLASTClust with a threshold of identity 30%, training set and the independent test set were really independent which would avoid the overfitting of prediction performance caused by high similar data. The datasets can be downloaded at https://sourceforge.net/p/praspkg.

2.2. Features

2.2.1. Position-based features

2.2.1.1. K nearest neighbors (KNN) score. In order to make use of cluster information of local sequence fragments for predicting amidation sites, here we found a query sequence fragment's K nearest neighbors in both positive and negative datasets according to sequence similarity by the following algorithm:

(1) For two local sequence fragments s_1 and s_2, when the window size is $2n+1$, the distance $D(s_1,s_2)$ between s_1 and s_2 is defined as

$$D(s_1, s_2) = 1 - \frac{\sum_{i=-n}^{i=n} sim(s_1(i), s_2(i))}{2n+1}$$

where *sim* as the amino acid similarity matrix, $s_1(i)$, $s_2(i)$ are derived from the BLOSUM62 substitution matrix (Henikoff and Henikoff, 1992) as $sim(a,b) = (M(a,b) - min(M))/(max(M) - min(M))$, where a and b are two amino acids, M is the substitution matrix and $max(M)$,$min(M)$ represent the largest, smallest number in the matrix respectively.

(2) The corresponding KNN feature is then extracted as follows: (i) Form a set of neighbors, known as the comparison set, by combining the positive and negative samples of the training set; (ii) Calculate distances from the query sequences to the samples in comparison set;(iii) Sort the neighbors by the distances and choose the K nearest neighbors; (iv) Calculate the percentage of positive neighbors in its K nearest neighbors as the KNN score.

In this paper, we chose K as 0.25%, 0.5% and 1% of the size of training set and extracted three KNN features – $KNN_{0.25\%}$, $KNN_{0.5\%}$ and $KNN_{1\%}$ for predicting amidation sites.

2.2.1.2. Terminal indicator. When the sample site is near to the sequence terminus, it's extremely possible that the number of amino acids in upstream or downstream is smaller than 10 (the window size is 21 here). In this situation, we added the extra amino acid 'X' to the empty positions and attributed the terminal feature to the sample fragment to discern if the site is near the terminus. We defined the terminal indicator where the value 0 represents that the query site is far away from the terminus and does not need to be appended, and the value 1 represents that, the query site is close to the sequence terminus and indeed needs extra 'X' to fill up the deficiency of fragments of 21 amino acids.

2.3. Physicochemical and biochemical properties features

AAindex (Kawashima et al., 1999) database provides numerical indices that describe various physicochemical and biochemical properties of amino acids. Furthermore, using multivariate statistical analyses, the high-dimensional attribute data in AAindex are summarized by five multidimensional patterns of attribute covariation that reflect polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge (Atchley et al., 2005). Then the amino acids can be encoded according to values associated with each summarized physicochemical property.

2.4. Predicted structure-based features

Here we extracted and coded 3 sub-types to represent predicted structure-based features including predicted secondary structure (SS), predicted disorder scores, accessible surface area (ASA) calculated from PSIPRED (Jones, 1999), DISOPRED (Ward et al., 2004), Scratch (Cheng et al., 2005) respectively.

2.5. Evolutionary information features

PSSM conservation score was used in combination with Gain/Loss together to reflect the evolutionary information of a protein sequence. The gain/loss of amino acids during evolution was calculated based on the normalized differences between the substitution numbers creating and removing the amino acid. And the corresponding list is available and each amino acid could be encoded according to values in the list (Jordan et al., 2005). In detail, the asymptotic frequencies of the gainers (losers), to be